The smoking chain: friendship networks, education, social background and adolescent smoking behavior in the Netherlands

Huisman, C.

Publication date
2013

2
Data and Methods

2.1 Introduction

This study uses probability sample survey data to make generalizable claims and longitudinal complex network[1] data to provide a clear understanding of the micro-level mechanisms of the effect of the interplay between secondary school friendship networks, social background characteristics, and education on adolescent smoking behavior. The reason for this combination is that, on the one hand, survey research that uses probability samples is needed to make generalizable claims about social reality. On the other hand, social reality exists of individuals and their relationships and creates a reality *sui generis*. Or, as Emile Durkheim puts it, '[S]ociety is not a mere sum of individuals. Rather, the system formed by their association represents a specific reality which has its own characteristics' (1966 [1895], p. 103). In contrast with probability sample-based survey research, which draws upon the independence assumption[2], complex network data and network analysis are more appropriate to address relationships.

The analytical strategy of this study consists of examining similar questions with cross-sectional data and longitudinal complex network data. This chapter first discusses how the data analyzed in this book were collected. Second, two of the methods used in this book are explained in more detail: community detection and actor-based models for longitu-

[1] A "complex network" means that the topological features of the network are non-trivial and cannot be reduced to those of a random network.

[2] The independence assumption refers to the independence of residuals (or errors). This means that the residuals (or errors) are distributed randomly and are not influenced by each other, such that they do not correlate.

dinal network analysis. Third, the analytical approach of this study is presented.

2.2 Sampling Procedure and Data Collection

2.2.1 Dutch National School Survey on Substance Use 2007

The data used in Chapters Three and Four come from the Dutch National School Survey on Substance Use 2007 (DNSSSU) (in Dutch, Peilstationsonderzoek 2007), a cross-sectional study of eleven- to eighteen-year-old elementary and secondary school students conducted by the Trimbos Institute. The sample consists of schools and classes within schools. First, the schools were randomly sampled, and then classes were randomly sampled within these schools. In total, 7415 students were interviewed (the response rate was 55 percent). The parents of 4119 students were also interviewed. Compared to non-responding parents, the parents who returned the questionnaire had younger children (t=-8.27, p<.001), were less likely to have smoked in their lives (t=-10.38, p<.001), were less likely to have smoked in the last month (t=-9.50, p<0,001), and were more likely to have attended higher educational levels (t=7.18, p<.001). No gender differences between children with non-responding and responding parents were found (t=1.31, p=.91). Because of differences in the duration of the different school types, only respondents from grades one to four were included to prevent an overrepresentation of intermediate general education (HAVO) and academic preparatory education (VWO) students. After deleting respondents with missing values, the number of cases was 3984 within 147 schools. For a more detailed description of the sampling and data collection procedure, see Monshouwer et al. (2008).

2.2.2 Longitudinal Network Data Dutch Adolescents (LNDA)

The data used in Chapters Three, Five, Six and Seven are longitudinal panel network data, referred to as the Longitudinal Network data on Dutch Adolescents, or LNDA. These data were especially collected for this study and are therefore tailor-made for the questions addressed.

The research question establishes the criteria for the data. First, given the interest in influence vs. selection, the questions require an examination of changes in smoking behavior among secondary school students over a period of time. The DNSSSU 2003 shows that the largest increase in previous month smoking prevalence (11.7 percent to 20.3 percent) among Dutch secondary school students is between the ages of thirteen and fourteen (Monshouwer, et al., 2004). Based on this information, second-graders were chosen as the prime research category.

Second, initially, an important part of this study concerns how school type and school-type composition relate to smoking. As discussed in the introduction, one of the questions this study addresses is whether school-type composition (in the sense of different school types located at one location vs. one school type at one location) affects smoking behavior. Therefore, this study required at least one school that housed the preparatory vocational school type, the intermediate general school type and university preparatory school type at the same location and one school that housed only one school at one location, preferably the lowest school type. This criterion resulted in the collection of data at one school that housed all three types, one school that housed only intermediate general education and university preparatory secondary education, and three schools that housed only the preparatory vocational education school type.

Third, there were several practical limitations to the project. Because of financial and time limitations, I had to interview all classes (44 in all) myself on two occasions with a seven-month interval. I chose to conduct these interviews because prior to beginning my PhD project, I worked as a drug and alcohol prevention educator in the province of Noord-Holland for more than five years (1999-2005). During that period, I educated thousands of secondary school students across different school types and learned skill and practice are needed to oversee and manage the (sometimes intense and chaotic) group dynamics of the secondary school classroom, especially in the lower school types. This experience convinced me that it was unwise to trust the important step of data collection to unskilled and inexperienced assistants, as is often the case in projects of this nature.

All school organizations[3] (N=19) in the province of Noord-Holland that met the above-mentioned criterion of school type composition were contacted by letter and phone. This contact resulted in ten positive responses. The school organizations that declined to cooperate gave two reasons: they were either too busy or were already involved in another research project. Due to the financial limitations of the project, five schools were included in the final sample. These schools were selected on basis of their optimal fit with the initial criterion of variation in school-type composition. The school that houses three different school types (approximately 1400 students) at one location is located in a

[3] One school organization can contain up to six schools.

municipality with a population of approximately 35,000 inhabitants. The other four schools, which are all part of a larger school organization (approximately 4500 students), are located in a municipality with a population of approximately 80,000 inhabitants. The first wave of data collection was conducted in October 2008, and the second wave was conducted in May and June 2009. Table 2.1 shows the numbers and percentages of second-grade students per school and school type in this sample. School One, with 156 students in the second grade, houses an intermediate general school type, an intermediate general/academic preparatory mixed school type, and an academic preparatory school type. School Two has 293 students in the second grade and houses all three school types. School Three has 104 students in the second grade and only offers the preparatory vocational school types. School Four also offers only preparatory vocational education and has 275 students in the second grade. School Five also offers only preparatory vocational education and has 133 students in the second grade.

Table 2.1 *Numbers and percentages of second-grade students per school and school type in the LNDA*

	Preparatory vocational				**Intermediate general**	**Intermediate general/preparatory academic**	**Academic preparatory**	**Total**	
School	**Praktijk-onderwijs**	**VMBO /basis**	**VMBO /kader**	**VMBO/ theoretisch**	**HAVO**	**HAVO/VWO**	**VWO**	**N**	**%**
1	0	0	0	0	28	106	22	156	16.2
2	0	0	0	82	118	0	93	293	30.5
3	51	35	18	0	0	0	0	104	10.8
4	0	48	75	152	0	0	0	275	28.6
5	0	15	43	75	0	0	0	133	13.8
Total	51	98	136	309	146	106	115	961	
%	5.3	10.2	14.2	32.2	15.2	11.0	12.0		100

Data collection. In collaboration with the schools, parents were informed by letter about the research project and were given the opportunity to refuse their child's participation. Only the parents of one student refused cooperation. In total, 44 classes were interviewed by means of a standardized questionnaire. The questionnaire in Wave One consisted of 45 questions, and Wave Two consisted of 34 questions. In both waves, all students were able to complete the questionnaire within the maximum time span of 45 minutes[4]. The students were told that the questionnaires would be treated confidentially and that they could refuse to participate. Only one student refused.

Social network analyses is sensitive to missing data (Burt, 1987). It is important to have the most complete data possible on the network and the people within the network to make correct estimations. The statistical techniques used in this study to analyze the data, which will be discussed in more detail later in this chapter, can manage missing values, to a certain degree (M. Huisman & Snijders, 2003; M. Huisman & Steglich, 2008). A value of up to 10 percent missing is acceptable, whereas more than 20 percent is problematic. Table 2.2 shows the absence of students in both waves. The total absence in Wave One was 8.8 percent, and the total absence in Wave Two was 11 percent. This is within the acceptable margin.

Table 2.2 *Absence of students in Wave One and Wave Two by school*

	School 1		School 2		School 3		School 4		School 5		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
Absent Wave One	6	3.8	15	5.12	11	10.58	24	8.73	12	9.02	68	7.1
Absent Wave Two	7	4.5	27	9.22	12	11.54	34	12.36	9	6.77	89	9.3
Absent Waves One and Two	2	1.3	5	1.71	4	3.85	3	1.09	2	1.50	16	1.7
Total	156	100	293	100	104	100	275	100	133	100	961	100

[4] The questionnaires were tested at other schools to determine whether 45 minutes (the average amount of time in one school period) was a reasonable amount of time to complete the questionnaire.

2.3 Community Detection

Chapter Three examines four descriptive questions on whether and how secondary school students cluster around the behavioral focus of smoking and how network clustering relates to secondary school compositional characteristics, such as school type and class, for which it uses so-called community detection. Network clustering refers to the notion that within a network, some subsets of nodes tend to connect more densely with each other than nodes across clusters. In the network literature, these denser clusters are referred to as communities (Reichardt & Bornholdt, 2006). With a community detection algorithm, clusters can be identified by means of a modularity approach. In a modularity approach, the actual network is compared to its randomized counterpart with the same degree of distribution. Communities (or groups) have relatively higher densities and are separated by parts of the network that have sparser densities compared to what is expected randomly. The algorithm maximizes the differences between actual and expected densities. Modularity maximization, introduced by Newman and Girvan, has become a widely used approach for community detection (Fortunato, 2010). It yields a network partitioning and a modularity score between 0 and 1, which is intuitively easy to interpret. A low score indicates that the number of within-community edges does not differ significantly from what one would expect at random. High values indicate strong clustering. In practice, values that indicate clustering fall between 0.3 and 0.7 (Newman & Girvan, 2004, p. 69).

The reason for using this technique is that social clustering occurs around foci. Foci can be "(…) persons, places, social positions, activities, and groups" (Feld 1981 p.1016), such as school classes. A core aspect of Feld's argument is that internal network structures and processes do not explain why people meet in the first place. Social structural factors precede network effects. Often, people are forced to interact with each other because of their institutional and/or physical context.

2.4 Actor-Based Models for Longitudinal Network Analysis

Chapters Five, Six and Seven all address the question of how to explain similarity in smoking behavior. As discussed in the introduction, an important issue addressed in this study is whether similarity in smoking behavior is due to friendship selection on similar smoking behavior or due to friends' influence. To refer to the tendency of friends to be similar in terms of salient individual attributes such as smoking behavior, Fararo and Sunshine (1964) coined the term *homogeneity bias*, or, in statistical terms, network autocorrelation(Steglich, Snijders, & Pearson, 2010). To address the question of what network processes lead to network autocorrelation, actor-based models for longitudinal network analysis developed by Snijders et al. (Snijders, 2001; Snijders & Baerveldt, 2003; Snijders & Duijn, 1997; Snijders, Steglich, & Schweinberger, 2007; Snijders, Steglich, & van de Bunt, 2010) and Steglich et al. (Steglich, et al., 2010; Steglich, Snijders, & West, 2006) are used. This technique can discriminate between selection and influence while accounting for other endogenous (or structural) network effects, such as reciprocity (the tendency of students to befriend someone who identifies them as a friend over a random other student in the network) and transitivity (the tendency to become a friend of a friend) as well as exogenous effects, such as school type, gender and various social background characteristics. More specifically, this approach can address three fundamental concerns in the investigation of selection and influence processes (Steglich, et al., 2010; Veenstra & Steglich, 2012).

First, there is the issue of the network dependence of the actors. A commonly used type of social network data is ego-centered network data. This type of data consists of self-reported information on the direct network neighborhood of randomly sampled individuals. This type of data makes it impossible to control for possible relations that were not selected by an individual respondent because it "precludes a meaningful assessment of selection processes. For adequately measuring selection effects, a meaningful approximation of the set of potential relational partners must be made, whose individual properties must be known irrespective of whether they actually become partners or not" (Steglich, et al., 2010, p. 339). A meaningful approximation of selection and influence processes can only be assessed if autocorrelation is allowed. This calls for a complete longitudinal network[5]. However, when observations are highly dependent, the use of common statistical methods is problematic because of the violation of the independence assumption. Second, to ensure that friendship formation is based on selection on similar smoking behavior or that similarity in smoking behavior between friends is due to social influence, it is necessary to be able to control for alternative mechanisms, such as similarity on the basis of

[5] In this study, complete network data refers to the network of secondary school students at one school. Preferably, data on the friendship network outside of the school and with students in other grades should also be included in the analysis. However, this is not practical in this study.

gender or structural network effects like reciprocity or transitivity. Third, behavior and network configurations change over time, and longitudinal network panel data consist of two or more snapshots at discrete points in time, resulting in incomplete observations. What happens between these two points in time is unknown. To address this issue, Simulation Investigation for Empirical Network Analysis (SIENA) software simulates the behavioral and network dynamics between these two points in time.

2.5 SIENA modeling[6]

SIENA can model over time a network that coevolves with individual behavior while statistically controlling for both. For SIENA models, two simplifications of the complexity of social life are made. First, for the simulated (unobserved) changes between two waves of observations, SIENA allows for a (possibly large) number of network and behavioral decisions made by individuals, but only one at a time. Continuous time is modeled as a series of discrete "ministeps," or intermediate points between two waves. At each ministep, *one* (randomly selected) actor can change a tie or his or her smoking behavior. This change modifies the options of the other actors in the focal actor's social environment. This simplification to one single action per ministep is not always valid and excludes coordinated actions that take place simultaneously (Snijders, van de Bunt and Steglich 2010). This limitation does not hinder our study of smoking, however, because to the best of our knowledge, no groups of adolescents decided collectively to begin or stop smoking at exactly the same moment. The number of ministeps in between waves, that SIENA estimates from the empirical data, thus depends on how many changes have taken place. The frequency of tie changes per unit of time between subsequent waves is expressed in SIENA by a *rate* (and a rate function). Second, a (randomly chosen) actor's decision at a given ministep depends only on the situation in which the decision is made, not on a longer past. This assumption makes it possible to represent the network-behavioral change, called "evolution," as a Markov chain.

SIENA handles decisions internally through two so-called *objective functions*, referring to objectives that people are supposed to have as shown in their tie formation and behavioral tendencies. The objective

[6] This description of SIENA modeling is taken from the appendix of the published paper based on Chapter Five, which is co-authored by Jeroen Bruggeman.

functions are linear in the effects specified by the researcher and are computed at each ministep in which a (hypothetical) tie is changed (established or dissolved) or behavior is changed. By optimizing the objective function under the effects and constraints discussed above, SIENA attempts to hone in on the network and behavior in the second wave through a trajectory of successive ministeps that starts begins in the first wave. SIENA simulates numerous scenarios (at least 1000 or more, determined by SIENA) and attempts to converge, such that the resulting model resembles the data in the second wave, assessed through the Method of Moments (Snijders, 2001). Convergence is reported in the output, and estimated parameters and their interpretation are similar to those of regression models: dividing a parameter by its standard error yields a t-value. Because SIENA simulates heuristically rather than computing exhaustively, a model provided after one round of simulations can differ slightly from a model obtained after another round depending how well the simulations converge. The parameters of the objective functions in the selection (i.e., tie formation/dissolution) and influence (i.e., behavior) parts of the model are un-standardized and cannot be directly compared with each other, but selection and influence are controlled for one another.

SIENA estimates several parameters. The selection and influence parts of SIENA display a *rate parameter* indicating the amount of change of the network and behavior between waves. The selection part also displays an *outdegree parameter*, signifying the people's inclination to have (friendship) ties at all. In most cases, it is negative. This may sound counterintuitive, but it means that the subjectively expected costs to establish a tie with a random individual who possesses no specific characteristic that makes him/her attractive outweigh the expected benefits. Furthermore, a positive outdegree means, in the long run, that the density of the network is constantly increasing, which does not happen in actuality.

The influence part of the model displays a *linear shape effect*. This indicates in a rectilinear way the tendency of an actor toward a particular score on a given behavioral variable at Wave Two depending on her score at Wave One (Snijders, et al., 2010). Alternatively, one can examine the effect in a curvilinear (*quadratic*) way. A positive quadratic tendency points to behavioral self-reinforcement through a positive feedback loop. Along these estimates, structural network (e.g., reciprocity) and attribute effects can be modeled. These effects will be discussed in more detail in Chapters Five, Six, and Seven.

2.6 Data Analytic Approach

The data analytic approach to Chapters Five, Six, and Seven is twofold. First, all the hypotheses are tested using two-level random intercept models and ego-centered network variables computed based on the LNDA. Random intercept models take account of the nested structure of the data (students nested within classes). The school level is not included as a separate level because it is collinear with the fixed effects of school type. The outcomes of the random intercept models must be interpreted with caution because the nature of the LNDA violates the independence assumption. This violation makes it impossible to generalize the findings to the entire population of Dutch secondary school students. The outcomes only tell us where to look with the SIENA models and provide information on the explained variance, something that is not possible with SIENA models. Second, to examine the effects of the influence of friendships and control for selection, SIENA models are used to test similar hypotheses. An important reason for modeling both random intercept regression models and SIENA models, which might be redundant at first sight, is that random intercept regression models are well suited to clarify general patterns of association that are more accessible to a broader reading public.

For the analyses in Chapters Five, Six, and Seven, the network data of five schools are merged into one dataset. In the data matrix, non-existing network tie possibilities are created. For example, a student at School Three can matrix-wise nominate a student of School Four. In reality, this is impossible because students can only nominate other students within the same school. Using a "structural zero" method, SIENA can consider the fact that this relationship between the students of School Three and Four is structurally not possible (Ripley, Snijders, & Preciado, 2011; Snijders, et al., 2010). However, Snijders et al. (2010, p. 51) point out that this structural zero method is only valid under the assumption that parameters are identical. When merging five schools, one should control for between-school variation using fixed effects models. However, a key variable in all chapters is school type; some of the schools in the sample have only one school type, whereas others have more than one but not all types (at one location). Because Cramer's V for school type by school location is 0.67, modeling both school type and school location leads to severe multicollinearity. For the same reason, a meta-analysis (Ripley, et al., 2011, p. 114) of the outcomes of five separate SIENA models, one for each school, would make it impossible to investigate the effect of variation in school type on smoking behavior. For this reason, this study will not control for school location.